ChinaXiv合作期刊
图书情报工作
第62卷 第10期 2018年5月

领域知识聚类性的动态演化分析*

■ 安宁[1] 滕广青[1] 白淑春[2] 毕强[3] 韩尚轩[1]

[1] 东北师范大学信息科学与技术学院 长春 130117 [2] 吉林大学图书馆 长春 130012

[3] 吉林大学管理学院 长春 130022

摘要：［**目的/意义**］探索领域知识发展过程中的聚类演化问题有助于揭示知识聚类的特征和规律，对于掌握知识生长演进过程中关联知识的聚集具有重要意义。［**方法/过程**］以复杂网络的思想为基础，基于标签邻接关系的发生值构建时间序列领域知识网络。即依据网络模体的理论，采用网络聚类系数的分析方法，对领域知识网络进行动态跟踪与分析；结合网络密度、特征路径长度、节点度值、封闭三元组等指标，从随机因素、度相关性、邻近关联3个方面对领域知识发展过程中的聚类演化现象进行分析。［**结果/结论**］研究结果表明：①领域知识在发展进程中始终保持较高的聚类性；②领域知识的聚类性同时包含随机性与结构性（非随机性）两方面因素；③领域知识聚类的动态状态在小世界网络和无标度网络之间摇摆演化；④领域知识的聚类状态在网络全局和局部节点之间表现出一定的差异性。

关键词：领域知识 知识网络 知识聚类 聚类系数

分类号：G255.76

DOI:10.13266/j.issn.0252-3116.2018.10.012

1 引言

长期以来，领域知识的群聚性问题一直是图书情报学界努力探索和揭示的问题。众多研究成果表明，在任何学科领域内部，其领域内的知识单元都不是以完全孤立与游离的状态存在，而是基于潜在的关联关系呈现出一定的团簇性与集群性。这种知识之间的关联关系随着领域知识的发展处于不断的变化中，从而使得知识的聚类也在发展中演化变迁。一方面，学科领域内的热点知识、核心知识会牵引关联知识不断聚集；另一方面，新知识的孕育和产生也会持续疏解这种聚集状态。因此，从时间序列的视角对领域知识聚类问题进行动态分析，把握和揭示领域知识发展过程中知识聚类的演化特征与规律，成为知识管理领域中亟需解决的问题。

有鉴于此，本研究以复杂网络理论为指导，基于社会化标注系统的当期发生值构建时间序列领域知识网络。采用网络分析中聚类系数分析方法，对领域知识发展演进过程中知识网络的聚类系数进行跟踪与分析。并结合知识网络的密度、特征路径长度、节点度值、封闭三元组、2跳路径等指标，对处于发展进程中的领域知识聚类问题展开研究，以期对领域知识聚类的演化状态及其规律做出有益的探索。

2 研究综述

最早把网络思维引入图书情报学研究领域的当属E. Garfield[1]和D. J. S. Price[2]。二人在20世纪50、60年代分别在《科学》（Science）杂志上发表论文，基于科学论文的引用关系构建引文知识网络，从网络思维的视角对科学知识的继承与发扬问题展开研究。随着20世纪末网络科学（Network Science）[3]的复兴，知识之间的团簇性、群聚性问题在网络科学的视角下得以重新诠释。知识节点之间的聚集程度使知识网络呈现复杂的拓扑结构，国内外许多学者开始关注基于网络分析的领域知识聚类问题。N. Shibata等[4]基于SCI

* 本文系国家自然科学基金面上项目“基于网络结构演化的Folksonomy模式中社群知识组织与知识涌现研究”（项目编号：71473035）研究成果之一。

作者简介：安宁（ORCID：0000-0002-9579-0150），硕士研究生；滕广青（ORCID：0000-0002-1053-0959），教授，博士生导师；白淑春（ORCID：0000-0001-6017-7156），副研究员；毕强（ORCID：0000-0001-7381-4986），教授，博士生导师；韩尚轩（ORCID：0000-0001-0962-3218）博士研究生，通讯作者，E-mail：hansx2017@sina.com。

收稿日期：2017-10-24 修回日期：2018-01-28 本文起止页码：85-93 本文责任编辑：徐健

和 SSCI 的引用数据，对不同类型的引文网络进行对比研究，发现在直接引文网络中，其聚类系数最大，表明通过直接引用连接的论文内容相似度最大，并且由于核心文献包含在最大网络组件中，其缺失的风险最小。李亚婷和马费成[5]基于 Folksonomy 知识组织模式，构建社会化标签共现网络，通过对该标签网络的计算分析，发现标签知识网络聚类系数的高位性（C = 0.816）。胡昌平和陈果[6]则将以三元闭包为基础的聚类结构用于关键词知识网络的层次结构分析，通过基于聚类结构的子层融合和大层区分，分析关键词知识网络中节点的微观关联结构，发现基于三元闭包的聚类结构能够有效揭示知识网络微观单元的多样性。

随着研究工作的深入，近年来知识网络的相关研究逐渐进入最具有挑战性的动态分析层面。M. E. J. Newman[7]根据数据库中的书目信息，对物理和生物领域中的合作网络的时间演变进行了实证研究。研究中通过前后时间窗口合作网络的对比分析，探测该网络是如何发生改变的，从而揭示出增长网络中的聚类与优先连接的规律与模式。J. Makani 和 L. Spiteri[8]通过对标签知识网络中标签增长、标签重用以及标签歧视 3 个指标的测度，发现具有独特性的标签数量随着时间的推移在稳步减少，反应社群知识的标签词汇的领域稳定性增强。W. Liu 等[9]利用美国物理学会的出版物数据集，通过逐年建立书目耦合网络（BCN），识别出代表不同研究领域的聚类，以冲积图的形式将物理研究中长期的知识演化进行可视化呈现，探索新知识是如何在旧知识的基础之上建立的。研究结果表明，大多领域的知识聚类都经历了较弱的波普尔混合，很少有领域是孤立或者经历过强烈的混合。刘向等[10]通过引入度择优连接和时间优先连接探测科学知识的继承与更新过程。其中度择优机制保证了对重要知识的连接，而时间优先连接机制则促成对最新知识的接受和知识的更新。研究指出，度择优导致了科学知识网络中马太效应现象的出现，虽为知识学习提供了方便但妨碍了知识的更新换代和新知识的脱颖而出，具有全局性的影响；而时间优先连接则具有局部影响的后发优势，在一定程度上平抑度择优所导致的马太效应的负面影响，这两种机制的结合形成了知识演化在研究基础与研究前沿之间的平衡。滕广青[11]从标签间的关联关系出发对领域知识网络中紧密型知识凝聚子群的发展过程进行时间序列的动态跟踪与分析。研究发现紧密型领域知识凝聚子群数量的波动与凝聚子群自身的扩张、衰减、派生、融合的演化过程有关；随着领域知识的发展，知识子群之间的交叠密度上升并基于交叠关系聚集成更大的知识群落。祝娜和王芳[12]从主题关联的角度入手，以 3D 打印领域为例，基于 LDA 识别出科技创新主题并进行分阶段细化分析，探测主题聚类内部与外部的关联强度。研究结果表明，在知识演化路径各阶段，新主题出现时必然携带新主题词出现，而一些主题的萎缩消亡必然导致相关主题词的萎缩消亡。

综上所述，随着网络分析理论与方法的日渐成熟，以网络思维对各类知识网络展开研究已经得到学术界的普遍认可。其中关于知识网络聚类问题的相关研究也取得了较为丰富的成果，甚至近年来的一些研究工作已经从静态分析发展到动态研究。然而，在领域知识的发展进程中，总是伴随着知识的生长、衰退、衍生、融合等现象发生。基于累计数据的动态分析（文献[13]、[10－11]等）侧重于对前序状态继承的生长性，基于发生值的分析（文献[8－9]、[12]等）则聚焦于知识演进变迁中的老化与创新。同时考虑到社会化标注系统更强的时效性，以及当前学术界对基于社会化标注系统构建知识网络的认可（文献[8]、[11]、[13]等）。为了更突出地捕捉和把握知识发展进程中这些变化对知识聚类产生的影响，本研究基于发生值构建时间序列标签知识网络，从随机因素、度相关性、邻近关联 3 个视角，对领域知识发展演进过程中的聚类问题展开分析与研究。

3 知识聚类相关理论

任何一个学科领域内的知识单元之间都存在一定的关联性，这种知识之间的关联关系或者是直接关联或者是间接关联，从而使学科领域内的知识单元不再离散无序，而是形成一定程度上的知识聚类。在一个学科领域发展演进的动态过程中，新生的知识及其关联关系总是在已有知识与知识关联的基础上产生的。这种新生的知识与知识关联更加突出地反映该学科领域在特定时期内的知识生长、衰落、衍生、融合等变化。因此，研究中采用领域知识当期发生值为基础数据，采用网络分析的思想与方法，对领域知识动态演进过程中的知识聚类问题展开研究。

通常来讲，聚类实质上是一种对集合内的研究对象进行重新分类的过程，在知识网络中这种重新分类过程通常体现为知识凝聚子群的形成。而领域知识在发展演进过程中，知识节点及其关联关系随着时间的变化而不断改变，使得领域知识在时间序列上不断出

现分化与聚合的现象。在基于发生值的知识演化分析中,上一个时期的部分知识节点和关联关系在下一个时期被隐含继承的同时,也有部分知识节点和关联关系在下一个时期中消退湮灭,同时还有新的知识节点和关联关系在下一个时期中新生。对于领域知识发展演进的过程而言,这是一种微规则循环往复作用的体现,也是领域知识演化迭代的发展过程。网络分析能够对其中的聚类现象进行量化,从而实现针对真实知识网络的计算与测度。

研究中主要依据网络模体(network motif)的思想对知识网络的聚类问题展开研究。网络模体是网络的构成单位,是 R. Milo[14]研究团队在他们发表于《科学》(Science)杂志的关于复杂网络基本构造区块的研究成果中首先提出的。其中的封闭三元组是多类网络模体中最适合揭示聚类关系的基本构造区块。在此基础上,进一步采用 M. E. J. Newman[15]推荐的聚类系数定义:网络中所有长度为 2 的路径中闭合路径所占的比例。考虑到研究中所构建的为无向知识网路,因此选择的聚类系数表达公式如下:

$$C = \frac{TC \times 3}{TP} \qquad 公式(1)$$

公式(1)中,TC 为封闭三元组(聚类视角下的网络模体,简称聚类模体)数量,即长度为 2 的路径中的闭合路径数量;TP 为连通三元组数量,即所有长度为 2 的路径数量,包括闭合路径和非闭合路径。由于每个封闭三元组中都包含 3 个连通三元组,因此系数为 3。对于公式(1)的具体解析如图 1 所示:

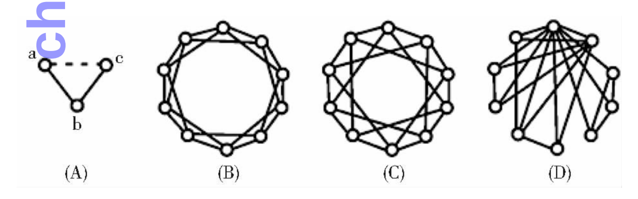


图 1 知识网络聚类系数解析

图 1 中,(A)为长度为 2 的路径 a-b-c,节点 a 与节点 c 拥有一个共同的邻居节点 b。在知识网络中通常也被理解为节点 a 与节点 c 有一个共同的关联知识节点 b。如果三元组{a, b, c}是封闭的(节点 a 与节点 c 之间也是实线连接),那么节点 a 与节点 c 也互为关联知识。因此 M. E. J. Newman 的聚类系数可以理解为知识网络中共同与某一知识节点直接关联的两个知识节点之间,相互直接关联的概率,即三点形成聚类模体的概率。也可以等价理解为与某一知识节点共同关联的两个知识节点,本身也直接关联的平均概率。图 1 中(B)、(C)、(D)是节点数量 N = 10,连线数量 L = 20 的同等规模、同等密度的网络。其中(C)为规则网络,由于(C)中并不存在任何封闭三元组(聚类模体),其聚类系数为 C = 0。另外,(B)为同样规模与密度的规则网络,(D)为同样规模与密度的非规则网络。由于(B)和(D)中包含一定数量的封闭三元组(聚类模体),因此二者拥有较高的聚类系数,其中(B)的聚类系数恒定与网络规模无关。由此可以发现,聚类系数与知识网络是否为规则网络无关,其中起到关键作用的是知识节点相互之间具备直接关联关系的聚类模体,即封闭三元组。

4 研究方法

4.1 研究数据

研究中,以社会书签和出版物共享网站 Bibsonomy 作为基础数据源展开研究。领域知识的选择范围可以是庞大的学科领域,也可以是精巧的主题领域,甚至是更细小的问题领域。由于近年来科学技术交叉融合的趋势愈发明显,领域知识也不再处于非此即彼的完全割裂状态,加之社会书签的开放性,许多原本似乎隔离的知识通过各类关联关系被纳入到相关的领域中,使得研究工作能够获得更广泛全面的视野。本研究以“folksonomy”作为目的标签,采用自主研发的网络爬虫工具,共抓取该领域相关文献 5 470 篇,时间跨度为 2006 - 2015 年。以自然年度作为时间刻度进行时段切割,将 2006 - 2015 时间区间划分为 10 个时间窗口(t0, t1, …, t9)。对各个时间窗口中的领域文献及其对应的标签进行统计,获得该领域各个时间窗口下文献与标签的相关基础数据如表 1 所示:

表 1 文献与标签数量的时间序列分布

时间窗口	文献数量(发生值)	标签数量(发生值)
t0	533	448
t1	865	901
t2	889	1010
t3	746	936
t4	625	862
t5	607	911
t6	349	690
t7	198	460
t8	273	689
t9	385	736

由于 A-L. Barabasi 和 R. Albert[16]在《科学》(*Science*)杂志上发表的关于网络标度涌现的文章中已经阐明“在复杂网络的发展过程中,总是伴随着节点的加

入、减少甚至是消失的现象”，同时考虑到本研究的视角主要聚焦于领域知识演进迭代过程中不同时期知识聚类的演化情况，因此为了更好地捕捉领域知识在生长演变过程中知识聚类的新生与衰落现象，表 1 中各个时间窗口的相关指标以当期发生值作为基础数据。

4.2 领域知识网络构建

基于社会化标注系统构建领域知识网络目前已经被学术界普遍认可和接受(参见文献[5]、[8]、[11]、[13])，研究中将根据抓取的数据构建领域知识网络。基于表 1 中的当期发生值数据构建标签邻接矩阵，以标签为网络节点，标签关联关系(同现关系)为网络连线，分别构建 t0 ~ t9 时间窗口的领域知识网络。其中，标签邻接矩阵采用二值矩阵。即标签 A 与标签 B 如果具备同现关系，则在邻接矩阵中记为 1，同时在知识网络中标签 A 与标签 B 所代表的知识节点之间由连线直接连接；若标签 C 与标签 D 不具备同现关系，则在邻接矩阵中记为 0，同时在知识网络中标签 C 与标签 D 所代表的知识节点之间则不存在直接连线。对各个时间窗口的领域知识网络中的节点与连线数量进行统计，所得结果如表 2 所示：

表 2 时间序列知识网络节点与连线数量统计

时间窗口	节点数量	连线数量
t0	448	2 899
t1	901	9 564
t2	1 010	7 129
t3	936	7 369
t4	862	5 069
t5	911	6 826
t6	690	5 650
t7	460	2 905
t8	689	6 375
t9	736	6 359

表 2 中，知识网络的节点数量即为标签数量(网络节点数量 = 当期标签数量)，连线数量即为标签同现关系数量(相同的标签同现关系不重复计数)。通过所构建的领域知识网络，能够将知识节点之间的关联关系呈现出来，并且从结构关系的视角揭示出知识聚类情况。由于时间序列的领域知识网络同时反映了领域知识发展演化的进程，因此基于标签和标签同现关系所构建的各个时间窗口的领域知识网络会随着领域知识发展过程中知识及其关联关系的繁荣与衰退，演化出其发展全程的多种形态。

4.3 领域知识网络聚类系数的提取

研究中采用上文以聚类模体为基础的聚类系数指标，对领域知识网络发展演进中的知识聚类问题进行分析与研究。按照不同的时间窗口计算提取领域知识网络的聚类系数。基于上文公式(1)计算获得各个时间窗口领域知识网络的聚类系数以及同等规模和密度的 E-R 随机网络的聚类系数如表 3 所示：

表 3 同等规模与密度的领域知识网络与随机网络聚类系数

时间窗口	Ck	网络密度	Cr
t0	0.292	0.028 9	0.031
t1	0.498	0.023 6	0.024
t2	0.193	0.014 0	0.015
t3	0.283	0.016 8	0.016
t4	0.181	0.013 7	0.014
t5	0.252	0.016 5	0.017
t6	0.425	0.023 7	0.023
t7	0.415	0.027 5	0.027
t8	0.56	0.026 9	0.026
t9	0.33	0.023 5	0.024

* 注：Ck 为知识网络聚类系数，Cr 为随机网络聚类系数

表 3 中的网络密度参数显示出 t0 ~ t9 时间窗口的领域知识网络密度均小于 0.03。网络密度的这一结果反映出，研究中基于真实数据构建的领域知识网络与目前发现总结的大多数真实网络一样，都属于一定程度的稀疏网络。聚类系数则反映网络的群聚程度。表 3 中知识网络的聚类系数为该领域当期真实数据计算所得，随机网络的聚类系数来自于同等规模和密度的 E-R 随机网络，用于研究中的对比参照。表 3 中知识网络的聚类系数同样不具备图 1 中(C)、(B)规则网络的特征(聚类系数为 0 或者规模无关性)。在表 3 数据的基础上，为了更全面地分析领域知识的聚类演化状态，研究中还将辅助以知识网络特征路径长度、封闭三元组数量等指标以及经典的统计分析技术，从随机因素、度相关性、邻近关联 3 个方面对领域知识演化迭代中的聚类演化情况展开研究。

5 研究结果

5.1 领域知识聚类的随机因素分析

从研究中抓取的原始数据(见表 1)来看，由于各个时间窗口的数据取值为当期的发生值，因此文献数量与标签数量的趋势与课题组此前的研究不同。文献数量与标签数量在时间轴的延展方向上并非是逐期递增的，而是基于当期实际发生数存在高低起伏的波动。同时结合表 2 中时间序列知识网络的相关数据可以发现，在同一时间窗口的截面下，知识间的关联关系数量

远远高于同期的知识节点数量。这一现象在反映了领域知识网络中知识之间的关联关系远比知识节点的数量更为丰富的同时,也从当期发生值的层面佐证了领域知识发展过程中知识之间逻辑关系的重要性[17],这也是动态网络稠化(densification)[18]的基础。

知识之间的逻辑关系在知识网络中表现为网络连线,连线的不同拓扑结构(封闭、连通等)能够对知识网络的聚类系数产生影响(见公式(1))。由于表3中知识网络聚类系数计算所用的数据集不包含前一时间窗口数据的叠加,因此并没有取得文献[19]中那样的跨越整个时间序列的极其显著的高位性。但是其取值范围与M. E. J. Newman[15]所总结的几种大规模知识网络的聚类系数的取值范围(0.088－0.45)基本相当。同时,将各个时间窗口的领域知识网络的聚类系数与同等规模、相同密度的E-R随机网络的聚类系数相比较可知,真实知识网络聚类系数的平均水平高出随机网络聚类系数平均水平15.8倍,即领域知识网络的聚类系数远高于E-R随机网络的聚类系数。这一现象反映出,基于当期发生值真实数据构建的领域知识网络,相比随机网络而言,在知识演化进程中仍然保持着较高的领域知识群聚性。由此可知,在领域知识的发展进程中,知识网络的群聚情况并非是随机网络一样的完全随机,而是存在着一定程度的非随机因素。此外,结合表3中网络密度的数据可以发现,E-R随机网络的聚类系数与网络密度之间具有显著的极强相关性(PEARSON相关系数Rce＝0.986 8),而真实的领域知识网络的聚类系数与网络密度之间的相关程度则明显低于前者(PEARSON相关系数Rck＝0.732 3)。显然,即使真实知识网络与随机网络保持了相同规模与相同密度,但是两类网络聚类系数与网络密度相关性的差异却进一步说明了真实知识网络中的关联关系存在一定程度的结构性(非随机性),这种结构性因素影响网络的聚类系数。

为了进一步检验领域知识网络发展演化过程中影响聚类系数的非随机因素,对各个时间窗口的领域知识网络的特征路径长度进行计算提取。所获得的领域知识网络的特征路径距离见表4。

根据表4中各个时间序窗口下领域知识网络中距离的分布情况可以发现,在整个时间区间内,该领域知识网络的路径距离中的1跳距离和4跳距离所占比例较小,2－3跳的距离占有显著的高比例。这一情况说明领域知识网络中任意2个知识节点在大多数情况下只需要2－3步可以实现连接。从知识网络全局范围

表4　时间序列领域知识网络的特征路径距离

时间窗口	1跳距离(%)	2跳距离(%)	3跳距离(%)	4跳距离(%)	平均距离
t0	2.90	49.50	43.90	2.30	2.500
t1	2.40	49.40	47.80	0.40	2.463
t2	1.40	44.70	53.30	0.70	2.532
t3	1.70	46.20	50.80	1.30	2.518
t4	1.40	39.20	56.20	3.20	2.613
t5	1.60	41.50	54.20	2.40	2.581
t6	2.40	45.80	49.20	2.60	2.521
t7	2.80	37.80	52.70	6.70	2.633
t8	2.70	46.70	48.70	1.90	2.497
t9	2.40	46.60	47.70	3.00	2.525

上来看,领域知识网络的平均路径长度在整个时间区间内始终保持在2.55±0.1范围之内。也就是说,该领域不同时期知识网络内部的节点之间平均只需要3步的距离就可彼此连通。结合此前领域知识网络的聚类系数远高于E-R随机网络的情况可以说明,较高的聚类系数与较短的特征路径长度符合D. J. Watts和S. H. Strogatz[20]当年在《自然》(*Nature*)杂志上提出的判定小世界网络的标准。因此,基于当期发生值的演化进程中的领域知识网络是处于随机网络和规则网络之间的小世界网络,而小世界网络独有的关联关系特征同时包含随机性与非随机性两方面的因素。

5.2　领域知识聚类的度相关分析

随机性与非随机性并存的具有小世界特征的领域知识网络,对网络聚类系数有着不同于一般随机网络的影响。T. G. Lewis[21]在对小世界网络的聚类系数的研究中发现,小世界网络的节点聚类系数比随机网络和无标度网络要高,并且倾向集中于具有中等度值的节点。然而在时间序列的动态演化过程中,时间维度的加入使得领域知识网络聚类系数的演化过程变得更为复杂。研究中将表3中的知识网络聚类系数与网络节点的度值结合进行分析。以网络中知识节点的聚类系数为纵轴,以网络中知识节点度值的对数为横轴,得到反映聚类系数与节点度值关系的散点图,结果见图2。

图2中t0时间窗口领域知识网络对应的散点图的密集中心(A区)位于中等偏低度值和中等偏高聚类系数的位置。这一特征几乎涵盖了时间序列的整个区间(t5时间窗口表现较弱),其中t1、t2、t3、t4、t6、t8、t9等时间窗口中该特征表现尤为明显。这一现象在一定程度上佐证了T. G. Lewis[2]在实验室条件下关于小世界网络聚类系数分布的研究结论。与此同时,t0时

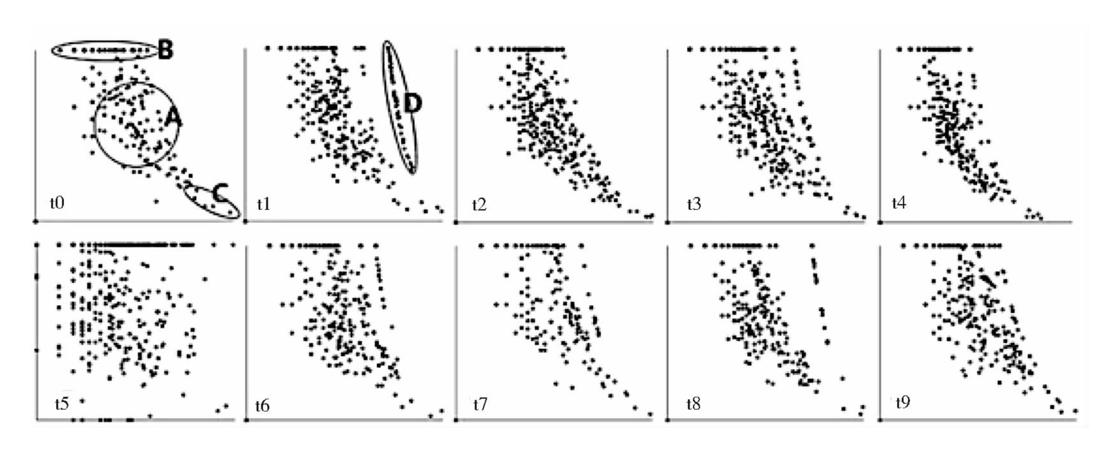


图 2 领域知识网络的节点聚类系数与度值关系

间窗口散点图顶部的最大聚类系数的点组成的近似横线(B 区)主要集中于中等偏低度值的区域,即中等偏低度值的知识节点有机会具有极高的聚类系数。这一现象更是小世界网络的节点聚类系数分布的极端表现。此外,t0 时间窗口散点区有一条比较鲜明的尾部(C 区)处于高度值和低聚类系数的区域,这一情况几乎延续了 t0 ~ t9 整个时间序列。同时,t0 ~ t9(t5 除外)时间窗口整个散点图总体上“头高尾低”的分布形态也近似地表明了高度值节点拥有低聚类系数,低度值节点具有较高聚类系数的特征。这一现象与实验室条件下无标度网络节点聚类系数的特征又基本吻合。因此,图 2 时间序列的数据分布在验证了领域知识网络具有小世界网络节点聚类系数特征的同时,也表现出无标度网络节点聚类系数的特征。

除此之外,图 2 中有 2 个现象需要特殊说明。其一,t1 时间窗口的领域知识网络对应的散点图中,点密集区的右侧的点形成一条近似于斜线的分布(D 区)。这一现象在 t1、t3、t6、t7、t8 等时间窗口中均有比较明显的表现。这一现象说明,在领域知识发展演进过程中,一部分拥有较高度值的知识节点也会随着领域知识的发展,在特定时期呈现出高低不同的聚类系数。即新产生的知识关联在一定程度上可以使较高度值知识节点的聚类系数提高。其二,t5 时间窗口领域知识网络对应的散点分布相对于其他时间窗口更为离散和随机,同时散点图顶部最大聚类系数的点组成的近似横线延伸跨度最长(包括低度值区域、中等度值区域和高度值区域)。这一现象说明领域知识发展演进过程中,知识聚类(有序)在个别特定时期仍然可能会表现出较强的随机性(无序)因素。

5.3 领域知识聚类的邻近关联分析

鉴于在此之前 A. Fronczak 等[22]的研究已经证明,随机网络中的聚类系数是由邻近网络的特征决定的,因此有必要从邻近关联的角度对领域知识的聚类进一步深入分析。在网络的拓扑结构中,聚类系数通常是对某节点的邻居节点之间建立联系的概率的评估。由于聚类系数是目标节点的两个邻居节点之间直接关联的概率,即节点 A、B 分别与节点 C 直接关联的情况下,节点 A、B 之间的直接关联的平均概率,因此当某节点的两个邻居节点一旦也具有邻接关系(直接关联)时,则形成了一种封闭的三元关系(聚类模体)。从这个意义上讲,可以理解网络的聚类系数事实上是对网络中这种聚类模体的度量。这种封闭的三元关系侧面反映了表面上距离为 2 的节点之间的仅通过一条连线就可相互关联。结合表 4 中对于领域知识网络中路径距离的分布情况可以看出,网络中 2 ~ 3 跳的路径占有明显的高比例。而其中的长度为 2 的路径说明邻近节点之间并没有形成封闭三元组,即节点之间需要 2 跳的距离才可以相互关联,也说明长度为 2 的路径事实上是对网络知识节点聚类的一种阻碍力量。为了更细致地分析领域知识网络中知识节点的聚类演化过程,进一步对领域知识网络中知识节点所形成的封闭三元组数量以及网络中长度为 2 的路径数量进行统计,其结果见表 5。

表 5 中的 2 跳路径为领域知识网络中尚未形成封闭三元组的一种连通三元组。根据时间序列下领域知识网络中封闭三元组、2 跳路径数量的统计结果,同时对照表 3 中领域知识网络聚类系数的统计结果可以发现,领域知识网络的封闭三元组数量与 2 跳路径数量的比值(TC/(TP-TC))和网络聚类系数之间存在较强的相关性,PEARSON 相关系数 Rcr = 0.973 9。由于基于真实数据的领域知识网络不可能具备实验室条件下网络规模不变等假设前提,因此不能简单地得出聚类

表 5　时间序列封闭三元组与 2 跳路径

时间窗口	TC	TP-TC	TC/(TP-TC)
t0	11 224	49 607	0.226
t1	112 232	198 496	0.565
t2	31 020	227 578	0.136
t3	44 012	202 172	0.218
t4	15 863	145 538	0.109
t5	32 918	171 993	0.191
t6	40 378	108 879	0.371
t7	12 683	39 401	0.322
t8	69 014	109 487	0.630
t9	38 006	126 024	0.302

* 注:TC-封闭三元组数量,TP-连通三元组数量,TP-TC-2 跳路径数量

系数与封闭三元组呈正相关,或者与 2 跳路径呈负相关的结论。结合上文公式(1)可以将这种相关性总结为,封闭三元组数量与 2 跳路径数量之间的比值在很大程度上决定了网络聚类系数的变化趋势。

在分析过程中,进一步选取该领域中具体的知识节点作为分析对象,在局部细节层面上对领域知识网络中知识节点的聚类演化过程进行动态跟踪。研究中选取"ontology"知识节点展开局部聚类演化过程的跟踪与分析,统计该知识节点的聚类系数、包含该知识节点的封闭三元组、以该知识节点为中介的 2 跳路径数量,以及封闭三元组与 2 跳路径数量的比值。相关结果如表 6 所示:

表 6　"ontology"的节点聚类系数、封闭三元组及 2 跳路径

时间窗口	Cn	TC	TP-TC	TC/(TP-TC)
t0	0.198	2043	1937	1.055
t1	0.302	86226	9673	8.914
t2	0.129	3604	5315	0.678
t3	0.205	1248	1269	0.983
t4	0.216	1580	1341	1.178
t5	0.251	5234	2367	2.211
t6	0.357	1790	581	3.081
t7	1	1	0	∞
t8	1	84	0	∞
t9	0.278	1246	814	1.531

* 注:Cn-节点聚类系数,TC-封闭三元组数量,TP-连通三元组数量,TP-TC-2 跳路径数量

从表 6 中的数据可以看出,在领域知识发展进程中,"ontology"知识节点周围的聚类情况并不是固定不变的。围绕该知识节点的封闭三元数量以及 2 跳路径数量都会随着领域知识的演进产生一定的变化,并且二者的变化走向大致相同。这反映了在领域知识的发展演进过程中,随着知识关联关系老化与新生的不断更迭,知识节点周围在大多数情况下会有新的代表非直接关联的 2 跳路径产生,同时也形成大量直接关联的封闭三元组关系。总体而言,表 5 和表 6 的结果都反映了通常情况下,封闭三元组(聚类模体)与 2 跳路径(通过长度为 2 的路径关联)的比值越高,该知识节点的聚类系数往往越高。其中 t7、t8 时间窗口 2 跳路径数量为 0,说明该阶段"ontology"的邻居节点中不存在无法直接关联的知识节点,表现出"ontology"领域知识发展演化进程中的阶段性稳定与成熟。

然而,作为复杂系统演化的知识网络发展还有其特殊的一面。众多的连通三元组(包括封闭三元组和 2 跳路径)也只是有利于促成节点在网络中的高 hub(高度值)地位。在某些情况下,即使封闭三元组与 2 跳路径数量具有较高的比例,也并非就一定能够完全决定以该知识节点为中心的聚类系数的高位性。表 6 中 t1 时间窗口"ontology"知识节点的封闭三元组数量远多于 2 跳路径数量且二者比值很高(比例 =8.914),但是其众多的 2 跳路径数量还是在一定程度上制约了 t1 时间窗口中"ontology"知识节点的聚类系数的高位性。因此,严格地说,相对于个体节点的 2 跳路径而言(非直接关联),封闭三元组(聚类模体)仅仅是高聚类系数的必要条件而非充分条件。

6　结论与讨论

本研究基于社会化标注系统中标签同现关系构建时间序列领域知识网络,探索领域知识演进过程中知识聚类的状态与相关影响因素。通过对时间序列领域知识网络的随机因素、度相关性、邻近关联等方面的跟踪分析,揭示领域知识演进过程中知识聚类的演化模式及其背后的影响因素。综合上述对领域知识演进过程的时间序列动态跟踪与分析,初步可以得出以下结论。

(1)在领域知识的生长发展过程中,领域知识始终保持较高的聚类性。从表 3 中知识网络的密度可以发现,基于真实的当期发生值构建的领域知识网络是一种稀疏网络。但是与同样密度、同等规模的稀疏的随机网络相比,领域知识网络的聚类系数明显高于随机网络。并且这种较高的聚类状态在整个演进周期内保持。尽管研究中使用的真实数据与 D. J. Watts 等[20]在实验室条件下使用的仿真数据不同,但是却从知识聚类的层面同样验证了知识网络与完全规则网络和随机网络的差异。即基于真实当期发生值构建的领

域知识网络,在知识演进过程中始终保持较高的聚类性。

(2)领域知识的聚类性同时包含随机性与结构性(非随机性)两方面因素。现实中的科学研究总是建立在前人研究工作的基础上,后续时间窗口当期发生值表现出的知识关联显然也经过潜移默化的更新和迭代,相当于一定程度上的网络重连。加之领域知识网络在演进过程中表现出的较高的聚类性(见表 3),以及 PEARSON 相关系数由极强相关(0.8 - 1.0)到一般强相关(0.6 - 0.8)的落差(Rce = 0.986 8、Rck = 0.732 3),其中的结构性(非随机性)因素也被凸显出来。同时,尽管在大多数时刻结构性(非随机性)因素表现显著,但是随机因素并没有完全泯灭。图 2 中 t5 时间窗口的散点分布的离散状态就表现出较强的随机性。因此,动态演进过程中的领域知识的聚类性,同时包含结构性(非随机性)与随机性因素。

(3)领域知识聚类的动态状态在小世界网络和无标度网络之间摇摆演化。较高的聚类系数(见表 3)和较短的特征路径长度(见表 4)表明,基于真实数据发生值构建的领域知识网络是一种小世界网络。同时,高聚类系数主要集中于中等偏低度值区域(见图 2 中 A 区、B 区)的现象也进一步体现出领域知识的聚类状态符合小世界网络的特征。然而在度相关分析中还可以发现,在时间序列的众多窗口中,散点分布呈现出的不同明显程度的尾部(如 t0 时间窗口散点图右下角的 C 区)。这一现象说明,领域知识发展演进过程中,知识的聚类状态同时还表现出无标度网络的特征。演进过程中不同散点区域的不同显著程度进一步反映了知识聚类状态在小世界网络和无标度网络之间摇摆演化。

(4)领域知识的聚类状态在空间维度(全局与局部)内也表现出一定的差异性。通过时间序列分析,领域知识聚类状态在时间维度上的差异已经跃然纸上。在空间维度方面,就领域知识全局(知识网络全局)而言,网络聚类系数与封闭三元组和 2 跳路径数量的比值呈正相关(Rcr = 0.973 9)。然而局部个体节点的聚类系数在总体趋势上保持上述正相关关系的同时,也在其中个别时刻呈现出一定的差异性。表 6 中 t1 时间窗口的高比值并没有获得极其显著的高聚类。显然,对于局部聚类系数而言,这种正相关关系并非是绝对严格的。从统计学意义上来讲,这也是样本均值与样本个体之间差异的体现,即全局知识聚类与局部知识聚类之间的差异。

本研究专门针对领域知识的当期发生值对知识聚类情况展开时间序列研究,基于发生值内含的知识衰老与新生,发现和揭示领域知识聚类状态与特征的演化规律。发生值与累计值相比,能够更好地捕捉和体现知识的老化与创新对领域知识聚类产生的影响。现实社会中的社交网络、信息传播网络等大多数真实网络都会面临友谊关系断绝或结交、传播渠道阻塞或新建等问题。因此研究中发现的聚类特征与规律,借助其对累计值数据中关联或连接关系不消除假设的解除,也有同样助于社交网络、信息传播网络等具有衰退和新生因素的网络的聚类特征的揭示。研究中的局限主要在于发生值相对于累计值而言具有更大的跳跃性,在对前序状态的继承性方面体现得不如累计值显著。在未来的研究中,将采取更兼顾平滑与跳跃的视角,将累计值与发生值相结合展开研究,以期更全面准确地对领域知识发展演化进程中的模式与规律进行探索与揭示。

参考文献:

[1] GARFIELD E. Citation indexes for science: a new dimension in documentation through association of ideas[J]. Science, 1955, 122(3159): 108 - 111.

[2] PRICE D J de S. Networks of scientific papers[J]. Science, 1965, 149(3683): 510 - 515.

[3] BARABASI A-L. Network science[M]. Cambridge: Cambridge University Press, 2016: 20 - 41.

[4] SHIBATA N, KAJIKAWA Y, TAKEDA Y, et al. Comparative study on methods of detecting research fronts using different types of citation[J]. Journal of the association for information science and technology, 2009, 60(3): 571 - 580.

[5] 李亚婷, 马费成. 基于标签共现的社会网络分析研究[J]. 情报杂志, 2012, 31(7): 103 - 109.

[6] 胡昌平, 陈果. 层次视角下概念知识网络的三元关系形态研究[J]. 图书情报工作, 2014, 58(4): 11 - 16.

[7] NEWMAN M E J. Clustering and preferential attachment in growing networks[J]. Physical review E, 2001, 64(2): 025102.

[8] MAKANI J, SPITERI L. The dynamics of collaborative tagging: an analysis of tag vocabulary application in knowledge representation, discovery and retrieval[J]. Journal of information & knowledge management, 2010, 9(2): 93 - 103.

[9] LIU W, NANETTI A, CHEONG S A. Knowledge evolution in physics research: an analysis of bibliographic coupling networks[J]. PLoS ONE, 2017, 12(9): e0184821.

[10] 刘向, 马费成, 王晓光. 知识网络的结构及过程模型[J]. 系统工程理论与实践, 2013, 33(7): 1836 - 1844.

[11] 滕广青. Folksonomy 模式中紧密型领域知识群落动态演化研究[J]. 中国图书馆学报, 2016, 42(4): 51 - 63.

[12] 祝娜，王芳. 基于主题关联的知识演化路径识别研究——以3D 打印领域为例[J]. 图书情报工作, 2016, 60(5): 101－109.
[13] MA J. The sustainability and stabilization of tag vocabulary in CiteULike: an empirical study of collaborative tagging[J]. Online information review, 2012, 36(5): 655－674.
[14] MILO R, SHEN-ORR S, ITZKOVITZ S, et al. Network motifs: simple building blocks of complex networks[J]. Science, 2002, 298(5594): 824－827.
[15] NEWMAN M E J. 网络科学引论[M]. 郭世泽，陈哲，译. 北京：电子工业出版社, 2014: 126－130, 152－173.
[16] BARABASI A-L, ALBERT R. Emergence of scaling in random networks[J]. Science, 1999, 286(5439): 509－512.
[17] 滕广青，贺德方，彭洁，等. 基于网络中心性的领域知识动态演化研究[J]. 图书情报工作, 2016, 60(14): 128－134, 141.
[18] MCGLOHON M, AKOGLU L, FALOUTSOS C. Statistical properties of social networks[C] // AGGARWAL C C. Social Network Data Analytics. New York: Springer, 2011: 17－39.
[19] 滕广青，常志远，刘雅姝，等. Folksonomy 知识组织模式中领域知识动态演化规律研究[J]. 图书与情报,2016(4): 96－101, 82.
[20] WATTS D J, STROGATZ S H. Collective dynamics of 'small-world' networks[J]. Nature, 1998, 393(6684): 440－442.
[21] LEWIS T G. 网络科学:原理与应用[M]. 陈向阳，巨修练，等，译. 北京：机械工业出版社, 2011: 138－140.
[22] FRONCZAK A, HOLYST J A, JEDYNAK M, et al. Higher order clustering coefficients in Barabási-Albert networks[J]. Physica A: statistical mechanics and its applications, 2002, 316(1): 688－694.

作者贡献说明：

安宁:数据分析与论文撰写；
滕广青:设计研究方案,数据分析,论文撰写及修订；
白淑春:数据采集与数据分析；
毕强:提出研究思路,设计研究方案；
韩尚轩:数据分析与论文修订。

Dynamic Evolution Analysis on Domain Knowledge Clustering

An Ning[1] Teng Guangqing[1] Bai Shuchun[2] Bi Qiang[3] Han Shangxuan[1]
[1] School of Information Science and Technology, Northeast Normal University, Changchun 130117
[2] Library of Jilin University, Changchun 130012
[3] School of Management, Jilin University, Changchun 130022

Abstract: [**Purpose/significance**] Exploring the clustering evolution in the process of domain knowledge development can help to reveal the characteristics and rules of knowledge clustering, this is great significance to master the clustering rules of correlation knowledge in the development and evolution process. [**Method/process**] Based on the idea of complex network, this paper constructed the time series domain knowledge networks in accordance with the occurred-value of tags adjacency relation. That is, according to the network motif theory, this paper dynamically tracked and analyzed the domain knowledge networks by the analysis method of network clustering coefficient. Then, by combining with the network density, the characteristic path length, the node degree value, the triadic closure and other indicators, this article analyzed the clustering evolution in the process of domain knowledge development from random factors, degree correlation, and adjacent correlation. [**Result/conclusion**] The results show: ①Domain knowledge in the development process always keeps a higher clustering. ②The clustering of domain knowledge includes both randomness and structuration (non-randomness). ③The dynamic status of domain knowledge clustering evolves between small-world network and scale-free network waveringly. ④The clustering status of domain knowledge shows a certain difference between the whole network and local nodes.

Keywords: domain knowledge　knowledge network　knowledge clustering　clustering coefficient